Volume 75
Number 4
1988

Annals
of the
Missouri
Botanical
Garden

TOWARDS MOLECULAR GENETICS IN *CLARKIA*: GENE DUPLICATIONS AND MOLECULAR CHARACTERIZATION OF PGI GENES[1,2]

L. D. Gottlieb[3]

ABSTRACT

I review the phylogenetic implications of eight duplications of nuclear genes encoding isozymes in Clarkia *(Onagraceae). These include ADH, cytosolic PGI, and both plastid and cytosolic isozymes of PGM, 6PGD, and TPI. The PGI duplication has been studied intensively from biochemical and genetic standpoints. Recent results have identified two levels of regulation that operate in species with this duplication, one that reduces cytosolic PGI activity to the level characteristic of species without the duplication (dosage compensation), and the second that results in differential accumulation of the products of the duplicate genes. These factors appear to reduce the impact of the duplication on metabolic function. I also describe our recent cloning and sequencing of two genes encoding PGI obtained from a genomic DNA library of* C. unguiculata, *a species with the duplication. The two genes encode proteins of 548 and 543 amino acids, respectively, and their predicted amino acid sequences are 58% homologous. They show 65% homology to a previously published partial amino acid sequence of pig muscle PGI. Both genes lack introns. The two genes are the first nuclear genes sequenced in wild plants. They are being studied as part of a research program on gene evolution and the application of nuclear gene sequences for phylogenetic reconstruction in higher plants.*

Questions about phylogeny have the form, "Is A more closely related to B than to C?" For flowering plants, the best phylogenies are thought to take into account the "maximum number of attributes possible" (Davis & Heywood, 1967: 485), with evidence from morphology, cytology, chemistry, reproductive compatibility, and other fields somehow combined. However, accurate phylogenetic reconstruction is more often a goal than an achievement because of problems brought about

[1] *This and the following three papers comprise the proceedings of the Missouri Botanical Garden's 34th Annual Systematics Symposium—Macromolecular Approaches to Phylogeny. The symposium took place in St. Louis, Missouri on October 9 and 10, 1987.*

[2] *The molecular genetics results (library construction, gene cloning and sequencing) described in this report were obtained in my lab by Dr. R. C. Tait, Debbie Laudencia, and Byron Froman. The molecular genetics research was supported by National Science Foundation grant BSR 86-07054 and USDA 86-CRCR-1-2139.*

[3] *Department of Genetics, University of California, Davis, California 95616, U.S.A.*

by character convergence, functional and developmental correlations, and unequal rates of evolution in different lineages. The essential difficulty is that little or nothing is known about how genetic changes that affect developmental processes result in differences in character expression.

The consequence is that no present procedure can translate the extent of morphological divergence into a measure of the closeness of phylogenetic relationship. I believe the way out of this impasse is to utilize a new source of evidence to assess phylogenetic relationships. The data of morphology, the traditional source of information about phylogeny, should be viewed as relevant to studies of plant development.

There is good reason to believe that information derived directly or indirectly from the structure and sequence of protein and DNA can be used to settle many phylogenetic questions. Interestingly, in this context, the molecular data are self-sufficient in that their usefulness does not depend on concordance with other lines of phenotypic evidence. For example, certain types of changes, particularly duplications of nuclear genes encoding enzymes (Gottlieb & Weeden, 1979; Gottlieb, 1983; Odrzykoski & Gottlieb, 1984) and large inversions of the chloroplast genome (Jansen & Palmer, 1987) appear to occur only once within a lineage. Thus taxa that now possess them probably descended from a single common ancestor and can be considered monophyletic without regard to their present morphological and cytological divergence.

In addition to phylogenetic inferences made on the basis of unique genetic and molecular traits, cladograms based on shared derived mutations or the extent of overall similarity can be constructed by comparing nucleotide sequences of genes or the size pattern of fragments cut from homologous DNAs by restriction endonucleases. The increasing availability of molecular data suggests that biosystematics no longer has to be considered an "unending synthesis" (Constance, 1964).

Phylogenetic relationships can now be determined accurately and reliably at many taxonomic levels. When this is done, the phylogeny can be used as a framework to ask important questions in other areas of biology. For example, how the attributes of species reflect both genetic legacy and selected and other changes since their origins, how genetic changes lead to specific modifications of ontogeny that result in new characters, and how and whether these new traits facilitate adaptation to different habitats. From this perspective, phylogeny can begin to inform both developmental and ecological analyses by providing evidence that A indeed derived from B and not from C.

In this paper, I review genetic and biochemical studies from my laboratory on gene duplications in *Clarkia* with emphasis on their phylogenetic implications. In addition, I describe very recent studies in which we have cloned and sequenced several genes encoding the glycolytic enzyme phosphoglucose isomerase from genomic libraries of *Clarkia* DNA. One purpose of these studies is to infer correct phylogenetic relationships in this well-studied plant genus. When the beginning and end points of species' genealogies are identified, we can ask about the steps in between.

Background

Previous to our studies and, indeed, making them appropriate, were the intensive investigations by Professor Harlan Lewis and his students and colleagues in the 1950s and 1960s (Lewis, 1953, 1962, 1973; Lewis & Lewis, 1955; Lewis & Raven, 1958). They correlated evidence from field studies, morphology, and a major program of hybridization and cytogenetical analysis. *Clarkia* was found to comprise at least 43 species, 33 being diploid. The diploid species were distinguished by substantial amounts of chromosomal repatterning in addition to aneuploidy. The extent of morphological divergence varied from a difference in a single character between some pairs of species to differences in entire suites of traits that might serve as evidence of generic distinction in other plant groups. The degree of morphological resemblance was frequently not concordant with the amount of chromosomal rearrangement. Nevertheless it was possible to discern meaningful phylogenetic patterns among the diploid species, and they were assigned to seven taxonomic sections (Lewis & Lewis, 1955). Allopolyploid species linked several sections so that as a whole the genus was considered a natural unit.

Lewis formulated an elegant model of speciation to account for these relationships. The critical features of this model included the following: (1) species were regarded as progenitor and derivative and not as siblings; (2) a new species differed from its parent by gross chromosomal rearrangements and sometimes by a change in basic number; (3) the speciation process was rapid and abrupt; (4) speciation was independent of the evolution of new adaptations and therefore was largely fortuitous; (5) speciation, in general, occurred at the xeric margin of the distribution of the parent species.

Lewis's model and his proposed examples of progenitor-derivative species made *Clarkia* appropriate for the first studies carried out in plants that applied electrophoretic analysis of enzymes to assess the amount of genetic divergence correlated with speciation (Gottlieb, 1973, 1974a). The rationale behind these studies has been reviewed by Crawford (1983, 1985) and by me (Gottlieb, 1977, 1981, 1986).

In addition to information about variation (presence, number, frequency) and divergence of alleles at loci coding enzymes, electrophoretic patterns provide evidence about the number of isozymes of particular enzymes and, thereby, the number of coding loci. As more and more species of *Clarkia* were examined, it became apparent that they sometimes differed among themselves or from other diploid seed plants in the number of isozymes of many particular enzymes. Subsequent genetic studies revealed that increased isozyme number resulted from duplications of the coding genes (Gottlieb, 1977; Gottlieb & Weeden, 1979; Pichersky & Gottlieb, 1983).

Examination of the number of isozymes in a broad array of higher plants, including conifers and angiosperms, showed that isozyme number was highly conserved and depended on the number of subcellular compartments in which a particular catalytic reaction occurred (Gottlieb, 1982). For example, in diploid plants, enzymes of glycolysis and the oxidative pentose phosphate pathway are encoded in the nucleus and are generally found as two isozymes, one located in the plastids and the other in the cytosol. When the number of isozymes within a particular compartment is more than one, it probably results from duplication of the structural gene or, in polyploid plants, from additive expression of the genes in the several constituent genomes. Since the conserved number of isozymes reflects the metabolic requirements of plant cells, a reduced number is not possible because it would be lethal. (Failure to observe bands of enzyme activity following electrophoresis of plant extracts should not be taken as evidence that the enzyme is not present in the extract, a common error in many surveys of electrophoretic variation in plants that report the absence of an expected enzyme band as a null allele.) The rules for recognizing duplicate isozymes, following electrophoresis of plant extracts, have been thoroughly described (Gottlieb, 1982). It is worth noting again that subcellular location furnishes the best criterion for recognizing the homology of isozymes from different species, and that the rules apply only to enzymes assayed with natural in vivo substrates. No regularities have been identified in the number of isozymes of enzymes such as esterases, phosphatases, and peroxidases that are generally assayed with artificial substrates.

Gene Duplication in *Clarkia*

The first duplicate isozyme discovered in *Clarkia* was that of alcohol dehydrogenase (ADH) in *C. franciscana* (Gottlieb, 1974b). Its absence from the closely related *C. amoena* and *C. rubicunda*, along with the very low genetic identity between *C. franciscana* and these species (Gottlieb, 1973), helped reject the hypothesis (Lewis & Raven, 1958) that *C. franciscana* was a recent derivative of *C. rubicunda*. The genetic evidence for duplication of ADH in *C. franciscana* was based on its exhibiting a true-breeding, three-banded electrophoretic pattern, whereas similar three-banded patterns in the related species resulted from heterozygosity at a single locus as evidenced by segregation patterns in progeny. Since *C. franciscana* did not display polymorphism for ADH, the duplication model was tested by making interspecific hybrids between it and *C. amoena*. The *C. amoena* plants used were homozygous at a single locus for an allele that encoded a slow ADH variant. The F_1 hybrids displayed a five-banded pattern that could only have resulted from the dimeric associations of three different polypeptides and, consequently, they must have possessed three genes (Gottlieb, 1974b). The ADH duplication was the second duplication of a gene encoding an enzyme discovered in plants. The first, in maize, was also an ADH (Schwartz & Endo, 1966).

Seven additional duplications of genes in *Clarkia* have since been described and, for each, the taxonomic distribution within the genus has been determined (Table 1). These duplications are cytosolic phosphoglucose isomerase (PGI) (Gottlieb, 1977; Gottlieb & Weeden, 1979), plastid and cytosolic triose phosphate isomerase (TPI) (Pichersky & Gottlieb, 1983), plastid and cytosolic 6-phosphogluconate dehydrogenase (6PGD) (Odrzykoski & Gottlieb, 1984), and plastid and cytosolic phosphoglucomutase (PGM) (Soltis et al., 1987). Detailed information about them is available in the individual reports. Five of the seven duplications (plastid and cytosolic 6PGD, plastid and cytosolic TPI, and plastid PGM) are present in species of all diploid sections of *Clarkia* (Table 1), suggesting they are at least as old as the genus. But only the duplicated plastid TPI was found in every species.

PLASTID AND CYTOSOLIC 6PGD

Four species of *Clarkia* appear to lack one or both 6PGD duplications (Odrzykoski & Gottlieb, 1984). *Clarkia rostrata* and *C. epilobioides* have a single plastid isozyme and a single cytosolic one and, consequently, lost both duplications. *Clarkia lewisii* and *C. cylindrica* have duplicated plastid 6PGDs but only a single cytosolic 6PGD (Table 1). The four species have been assigned to sect. *Peripetasma*, with the morphologically similar and crossable (Davis, 1970) *C. rostrata*, *C. lewisii*, and *C. cylindrica* to one subsection and the distinctive and highly self-pollinating *C. epilobioides* to a monotypic subsection (Lewis & Lewis, 1955). The close relationship of the former three species suggested that the loss of the duplicated cytosolic 6PGD occurred in their common ancestor and was subsequently followed in *C. rostrata* by an additional mutation or chromosomal deletion that silenced a duplicated gene encoding a plastid 6PGD. Since *C. epilobioides* also lacked both duplications, it seemed reasonable to suggest that it was closely related, although it was not possible to decide if the loss of its plastid 6PGD duplication was independent of the loss in *C. rostrata*.

The matter was settled by a restriction endonuclease analysis of chloroplast DNA carried out on all the species in this section, which revealed that *C. rostrata* and *C. epilobioides* were sister species and that *C. lewisii* and *C. cylindrica* comprised a second pair of sister species (Sytsma & Gottlieb, 1986a). The chloroplast DNA study also showed that the two pairs of species share a common ancestor well removed from the other species of the section. Thus, even though *C. rostrata* is not morphologically similar to *C. epilobioides* and was placed in a different subsection, the two species have a close genealogical relationship. Since this phylogenetic inference was based on evidence from both nuclear genes and chloroplast DNA, it is particularly strong.

PLASTID AND CYTOSOLIC TPI

Both TPI duplications appear to be present throughout *Clarkia* (Table 1), although some uncertainty remains in regard to the cytosolic TPI in sect. *Eucharidium* for which the genetic analysis is incomplete (Pichersky & Gottlieb, 1983). Electrophoretic studies of TPI have also been carried out on a number of species of other genera of Onagraceae to ascertain the taxonomic distribution of the duplications outside of *Clarkia*. Since sufficient (or appropriate) material was not available to conduct genetic analysis, three criteria had to be met to warrant the hypothesis that a given species possessed a TPI duplication. The minimum number of electromorphs per individual for each isozyme had to be at least three (TPI is dimeric), the multiple isozymes had to be located in the same subcellular compartment, and a side-by-side comparison of leaf and pollen extracts had to show the same number of cytosolic isozymes (the criteria are discussed in detail in Gottlieb, 1983). On the basis of satisfying all of these criteria (although sample sizes were very limited), the cytosolic TPI duplication was identified in five of the seven tribes of the family, including Jussiaeeae (*Ludwigia*), Fuchsieae (*Fuchsia*), Hauyeae (*Hauya*), Onagreae (*Clarkia*, *Heterogaura*, *Camissonia*, *Calylophus*, *Gongylocarpus*, and *Oenothera*), and Epilobieae (*Boisduvalia*) (Pichersky & Gottlieb, unpubl.). The presence of the duplication in both *Fuchsia* and *Ludwigia*, the two most ancient lineages in the family (Raven, 1979), suggests its great antiquity. In contrast, the plastid TPI duplication was not identified outside of *Clarkia* and must have arisen much more recently. Although these results should be regarded as exploratory, they point out the possibility that certain taxonomically widespread duplications may be useful to group genera (and eventually families) into monophyletic assemblages. However, since the time spans in these comparisons are great, it would be appropriate and necessary to validate the conclusions by examination of the nucleotide sequences of the duplicated genes.

PLASTID AND CYTOSOLIC PGM

In contrast to the situation in 6PGD in which the absence of duplicated genes could be assigned to some type of mutation in common ancestors of extant species, the loss of the plastid PGM duplication (Table 1) in *C. concinna* and in *C. lassenensis* (Soltis et al., 1987) must be regarded as independent events in lineages directly ancestral to these species but to no others, since the two species belong to distantly related sections of *Clarkia* (Lewis & Lewis, 1955).

The presence of the cytosolic PGM duplication in *C. arcuata* (sect. *Rhodanthos*) and in all species of sections *Godetia* and *Myxocarpa* (Table 1) is consistent with a taxonomic assignment previously made by Lewis & Lewis (1955). They proposed (p. 261) that sect. *Rhodanthos* (then designated sect. *Primigenia*) was "probably directly ancestral" to sect. *Godetia* and "perhaps" to sect. *Myxocarpa*. Within sect. *Rhodanthos*, the relevant lineage is now represented by *C. arcuata*

TABLE 1. *The phylogenetic distribution of duplicate isozymes in diploid species of* Clarkia. *The PGI data are from Gottlieb & Weeden (1979), the PGM data from Soltis et al. (1987), the 6PGD data from Odrzykoski & Gottlieb (1984), and the TPI data from Pichersky & Gottlieb (1983). The numeral 1 indicates the species has a single isozyme and the numeral 2 indicates duplicated isozymes. For each enzyme, plastid (Pl) and cytosolic (Cy) isozymes are indicated.*

Section Species	Isozyme Number						
	PGI	PGM		6PGD		TPI	
	Cy	Pl	Cy	Pl	Cy	Pl	Cy
Eucharidium							
C. breweri	2	2	1	2	2	2	?
C. concinna	2	1	1	2	2	2	?
Fibula							
C. bottae	2	2	1	2	2	2	2
Peripetasma							
C. cylindrica	2	2	1	2	1	2	2
C. lewisii	2	2	1	2	1	2	2
C. epilobioides	2	2	1	1	1	2	2
C. rostrata	1	2	1	1	1	2	2
C. biloba subsp. *australis*	2	2	1	2	2	2	2
C. dudleyana	2	2	1	2	2	2	2
C. lingulata	2	2	1	2	2	2	2
C. modesta	2	2	1	2	2	2	2
Heterogaura heterandra	2	2	1	2	2	?	2
Phaeostoma							
C. xantiana	2	2	1	2	2	2	2
C. unguiculata	2	2	1	2	2	2	2
Godetia							
C. imbricata	1	2	2	2	2	2	2
C. nitens	1	2	2	2	2	2	2
C. speciosa subsp. *polyantha*	1	2	2	2	2	2	2
C. williamsonii	1	2	2	2	2	2	2
Myxocarpa							
C. mildrediae	1	2	2	2	2	2	2
C. virgata	1	2	2	2	2	2	2
Rhodanthos							
C. arcuata	1	2	2	2	2	2	2
C. lassenensis	1	1	1	2	2	2	2
C. amoena subsp. *huntiana*	1	2	1	2	2	2	2
C. franciscana	1	2	1	2	2	2	2
C. rubicunda	1	2	1	2	2	2	2

which, together with *C. lassenensis*, was placed in a distinct subsection. The other subsection containing diploid species includes *C. amoena*, *C. rubicunda*, and *C. franciscana*, and it then would represent the lineage from which the other four sections of *Clarkia* (Table 1) eventually evolved. Alternatively, the cytosolic PGM duplication may have had independent origins in *C. arcuata* and sections *Godetia* and *Myxocarpa*. Sequence comparisons of the PGM genes will make it possible to distinguish these models.

Regardless of the outcome of such comparisons, the taxonomic distribution of the cytosolic PGM duplication is independent of the sectional phylogeny suggested (Lewis, 1980) following the discovery of the cytosolic PGI duplication (Gottlieb, 1977; Gottlieb & Weeden, 1979), since the two duplications are not present together in any species (Table 1). The PGM evidence suggests that the four sections that have the PGI duplication (Table 1) arose from the lineage within sect. *Rhodanthos* that also gave rise to *C. amoena*, *C. rubicunda*,

and *C. franciscana*. It is also an interesting possibility that since the two enzymes catalyze adjacent reactions in glycolysis and gluconeogenesis (PGI interconverts fructose-6-phosphate and glucose-6-phosphate, and PGM interconverts the latter and glucose-1-phosphate), there may be metabolic reasons that select against the occurrence of both duplicated enzymes in the same cytosol.

Overall, the genetic and biochemical evidence from the several gene duplications provides a remarkably consistent and coherent picture of the phylogenetic relationships within *Clarkia*. The evidence is also consistent with the recent discovery based on restriction endonuclease patterns in chloroplast DNA that the monotypic *Heterogaura heterandra* (Table 1) is actually a *Clarkia* and closely related to *C. dudleyana* (Sytsma & Gottlieb, 1986b).

PGI

The PGI duplication in *Clarkia* has been studied intensively because it was one of the first duplications identified that is present in some but not all species of a single genus. Thus, it is possible to compare duplicate PGI genes and their products with their nonduplicate homologues, and the comparisons can be done in species having a relatively similar genomic background. The example provides an unusual opportunity to examine the critical early stages of gene evolution and to test the general model that major changes in gene regulation, structure, and function cannot evolve without the availability of duplicate sequences.

The PGI duplication characterizes all of the species (except *C. rostrata*) in the morphologically advanced and diverse sections *Eucharidium*, *Fibula*, *Phaeostoma*, and *Peripetasma*, and is absent from sections *Godetia*, *Myxocarpa*, and *Rhodanthos* (Table 1). Consequently it identifies a specific branching point in the phylogeny of *Clarkia* and serves to group the former four sections into a monophyletic lineage (Gottlieb & Weeden, 1979; Lewis, 1980). The realignment was effected without having to move any species into or out of any section (Lewis, 1980).

Genetic studies revealed that the duplicate PGI genes assort independently (Gottlieb, 1977; Gottlieb & Weeden, 1979; Weeden & Gottlieb, 1979), which is thought to mean that they arose by a process involving overlapping reciprocal translocations or insertional translocations rather than unequal crossing-over. The relevant arguments were presented in Gottlieb (1983). Many other duplicate genes in plants also assort independently (Tanksley, 1987). The mode of origin is important for phylogenetic reconstructions because chromosomal rearrangement is much more likely than unequal crossing-over to occur only once for a particular chromosome segment in a particular linkage. Linkage relationships for the other duplications in *Clarkia* have not been studied in similar detail, although we do know that the duplicate genes encoding plastid TPIs and one of them and a cytosolic TPI gene also assort independently (Pichersky & Gottlieb, 1983).

A number of biochemical studies were carried out to determine how much and what type of divergence marked the duplicate PGI isozymes. Three results are noteworthy, one having to do with the molecular weight of PGI subunits and the other two with the evolution of regulatory factors that appear to modulate the expression of the duplicate PGI genes.

PGI subunits encoded by the duplicate genes have different apparent molecular weights (apparent because the values were obtained from their electrophoretic mobility on SDS gels), with PGI-2 being 60,400 and PGI-3 59,000, or values closely similar (Gottlieb & Higgins, 1984a). Species in sect. *Myxocarpa* that lack the duplication have PGI subunits with molecular weight of 60,400, and PGIs from sections *Godetia* and *Rhodanthos* weighed in at 59,000. The presence of two molecular weight forms in species with the duplication and each molecular weight form by itself in species without the duplication was unlikely to have come about by chance. The result suggested the novel possibility that the PGI locus in an ancestal *Clarkia* was translocated to different nonhomologous chromosomes, that the genes then accumulated mutational changes that encoded different molecular weight subunits, and that lines carrying the different chromosomes eventually hybridized with both PGI genes becoming segregated into a single genome by a process originally documented in maize (Burnham, 1962) involving overlapping reciprocal translocations. The scenario seems feasible for *Clarkia*, in which species are distinguished by gross amounts of chromosomal rearrangement, and which all have a self-compatible breeding system permitting chromosomal heterozygotes to be made homozygous and true-breeding by self-pollination. The merits of this speculation can be directly tested by comparing nucleotide sequences of PGI genes from species with and without the duplication (see below).

After it became apparent that the catalytic properties of the duplicate and nonduplicate PGIs were alike (Higgins & Gottlieb, 1984), studies turned to questions about increased gene dosage and whether

it caused increased levels of cytosolic PGI activity and protein. The PGI levels in clarkias with and without the duplication were assessed by immunological means using an antiserum specific to cytosolic PGI (i.e., one that does not cross-react with plastid PGI). The result was clear-cut. The two types of species had the same levels of cytosolic PGI activity and protein, suggesting that some form of regulation had evolved that "compensated" for the duplicated genes (Gottlieb & Higgins, 1984b). The activity level proved to be the same as that in a number of diploid vegetables, indicating that green plants generally maintain a similar PGI level. This finding provided an important rationale for the evolution of dosage compensation because it restored an activity level characteristic of typical diploid plants having a single cytosolic PGI. Thus a regulatory mechanism had evolved that reduced the impact of the duplication on metabolic function.

To determine whether the regulation operated via metabolic or genetic factors, a series of null activity mutants of each duplicate gene was induced by ethyl methanesulfonate (EMS) treatment of seedlings of *C. xantiana* (Jones et al., 1986). Metabolic factors would be implicated if lesions induced in either gene did not change PGI levels. In homozygous state, each mutant completely lacked the homodimer activity normally specified by the affected gene. The mutants were backcrossed to wildtype for five generations, making it possible to assign changes in PGI activity directly to the mutation and not to unknown factors in the background. Immunological analysis revealed that they reduced PGI activity in direct proportion to the normal contribution of each gene. The homozygous mutants at *Pgi2* reduced cytosolic PGI activity to 36% of wildtype, and the mutant at *Pgi3* to 64%. The effects of the mutations at the two loci were additive. Thus, $Pgi2^{null}\,2^{null}$, $Pgi3^{a}\,3^{null}$ plants synthesized in an F_2 progeny from experimental hybrids between the two mutants exhibited only 14% of wildtype activity. The double homozygous null was lethal. The results demonstrated that PGI activity in plants having the duplication is not directly regulated by metabolic factors, warranting the suggestion that the dosage compensation depends on factors that regulate the levels of transcription or translation (Jones et al., 1986). Since *Pgi3* contributes less than *Pgi2* to the total cytosolic PGI activity, the regulatory factors appear to operate to a greater extent on the former locus. Thus, two levels of regulation were identified, one that reduces cytosolic PGI activity in species with the duplication to the level characteristic of species without the duplication, and the second that results in differential accumulation of the products of the duplicate genes.

The genetic and biochemical analyses of PGI in *Clarkia* identified a number of interesting questions that can be answered only with evidence from the sequences of the coding genes. For example, in terms of phylogenetics, it is necessary to test the major hypothesis that the duplication had a unique origin, with the consequence that the four sections that possess it are monophyletic. A corollary hypothesis is that the origin of the duplication involved hybridization between lineages now represented by *Myxocarpa* (which has the higher molecular weight subunit) and *Godetia/Rhodanthos* (with the low molecular weight subunit). The hypotheses can be tested by comparing the sequences of duplicate and nonduplicate PGI genes. On the hypothesis, *Pgi2* from a species with the duplication should be similar to *Pgi* from *Myxocarpa*, and *Pgi3* from the species with the duplication should be similar to *Pgi* from *Godetia/Rhodanthos*. In other words, the duplicate genes should be more similar to genes from different species than they are to each other.

Other questions of interest in a context of evolutionary biology have to do with the extent of PGI sequence divergence in species with the duplication versus those without the duplication, the extent of polymorphism for PGI genes in natural populations of *Clarkia*, and the value of the sequences to demonstrate phylogenetic relationships outside of *Clarkia*, particularly among the diverse genera included in tribe Onagreae.

A different set of questions must be answered to explain how the cytosolic PGI level is reduced in species with the duplication to that characteristic of those without it, to determine the basis for the near 2:1 difference in PGI activities attributable to the duplicate genes, to learn why *Pgi2* encodes a higher molecular weight unit than *Pgi3*, and the nature of the mutations that eliminated PGI activity in the EMS-induced null mutants.

Clarkia PGI Gene Sequences

Headway on these questions can now be made because we have cloned and sequenced PGI genes from several *Clarkia* genomic libraries. Here I describe how these genes were obtained, evidence that they encode PGI, and their general structure. Detailed characterizations and sequences will be presented separately. To my knowledge, the *Clarkia* PGI genes are the first nuclear genes from wild plants that have been sequenced.

Our first genomic library was constructed with DNA isolated from seedlings of a horticultural strain

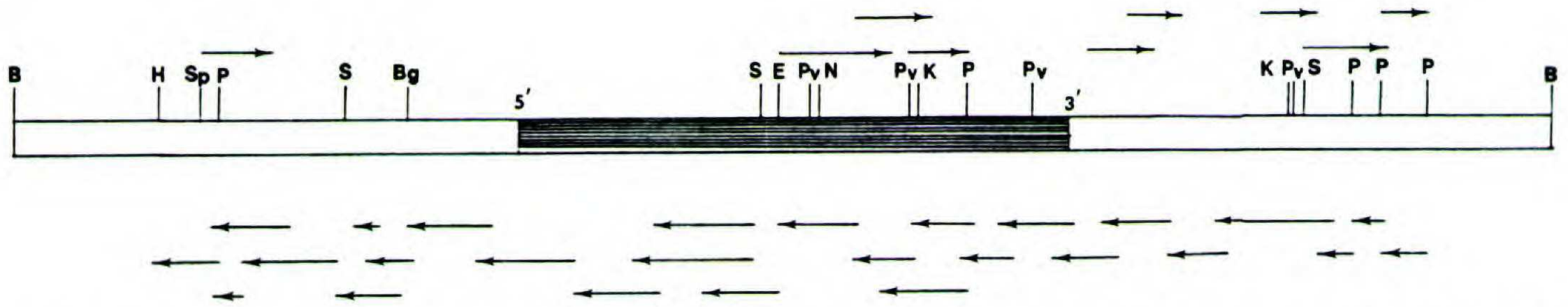

FIGURE 1. *Restriction map and sequencing strategy for the* Clarkia unguiculata *U2 gene which encodes PGI. The gene is present on a 4.45-kb* Bam*H1 fragment. Restriction sites shown are* Bam*H1 (B)*, Hpa*I (H)*, Sph *(Sp)*, Pst*I (P)*, Sal*I (S)*, Bg*1II (Bg)*, Eco*R1 (E)*, Pvu*II (Pv)*, Nco*I (N)*, and* Kpn*I (K)*. The arrows above and below the restriction map show the direction and extent of sequencing for the individual M13 subclones. 3.8 kb including the entire coding region was sequenced on one strand, and 1.6 kb on the complementary strand.*

of *C. unguiculata* (Northrup King "Clarkia Double Mixed Colors"), a species with the PGI duplication. Horticultural material was used because very large amounts of seed could be purchased, permitting us to fine-tune our techniques prior to studying natural populations. The DNA was extracted by a procedure modified from Fischer & Goldberg (1982) that yielded a nuclear pellet that, after lysis, provided high molecular weight DNA fragments (greater than 100 kb). The DNA was partially restricted with *Sau*3a and fragments between 15 and 23 kb obtained by fractionation on a sucrose gradient. After determining optimal ratios of chromosomal DNA and vector arms, the DNA fragments were cloned into the *Bam*H1 site of the lambda replacement vector Charon 35. The resulting library is estimated to contain 1.8×10^6 phage with 88% recombinants and represents about seven *Clarkia* genomes.

The library was screened at low stringency (51°C, 5 × SSPE) with an 800-bp DNA fragment of a yeast gene encoding PGI, kindly provided by a biotechnology firm. Since we expected low to very low homology between the yeast and *Clarkia* PGI sequences, the screening conditions were determined in a prior experiment in which the probe was hybridized on a Southern blot to genomic *C. unguiculata* DNA digested with several restriction enzymes. Two positive clones were obtained from the first 30,000 plaques examined. They were purified, and DNA prepared from each was restricted with several enzymes, subjected to agarose gel electrophoresis, and analyzed by Southern blots using the yeast PGI DNA fragment as probe. The two clones had inserts of 13.7 and 15 kb, which proved different. Hybridizing fragments of the former clone, designated U2, were cloned into M13mp10 and partially sequenced. The sequences showed homology to that of the yeast gene. A 4.45-kb *Bam*H1 fragment (Fig. 1) was then subcloned into pUC19 and deletion fragments constructed using the exonuclease III-S1 protocol of Henikoff (1984). One strand of 3.8 kb including the entire coding region was completely sequenced, and 1.6 kb was sequenced on the complementary strand by the dideoxy sequencing protocol (Messing, 1983). The U2 sequence revealed an uninterrupted open reading frame of 1,644 nucleotides encoding a protein of 548 amino acids.

TABLE 2. *Homology between predicted amino acid sequences from nucleotide sequences of U2 and U8, cloned from a genomic library of* Clarkia unguiculata *and amino acid sequences of cyanogen-bromide peptides purified from pig muscle PGI (Achari et al., 1981).*

Sequence	Number Identical Amino Acids / Total Number of Amino Acids	Homology
U2 vs. U8	319/548	58%
U2 vs. Pig	110/166	66%
U8 vs. Pig	108/166	65%
U2 vs. U8 (in sequences covered by Pig peptides)	89/165	54%

The identity of U2 was established by comparing its predicted amino acid sequence with the amino acid sequences of five cyanogen-bromide peptides obtained from pig muscle PGI (Achari et al., 1981). These are the only PGI sequences, protein or DNA, that are published for any organism. The five pig peptides identify a total of 166 amino acids, about 30% of the complete protein. The U2 gene encodes amino acids that are identical to those in pig PGI at 110 of these 166 residues, or 66% of the total (Table 2). A second PGI gene, called U8, also obtained from the *C. unguiculata* genomic library, using U2 as the probe, was found that contains the same sequence present in the 15-kb insert noted above. A similar isolation and sequencing strategy was used to characterize the U8 clone as was used for U2. U8 proved to have a 65% ho-

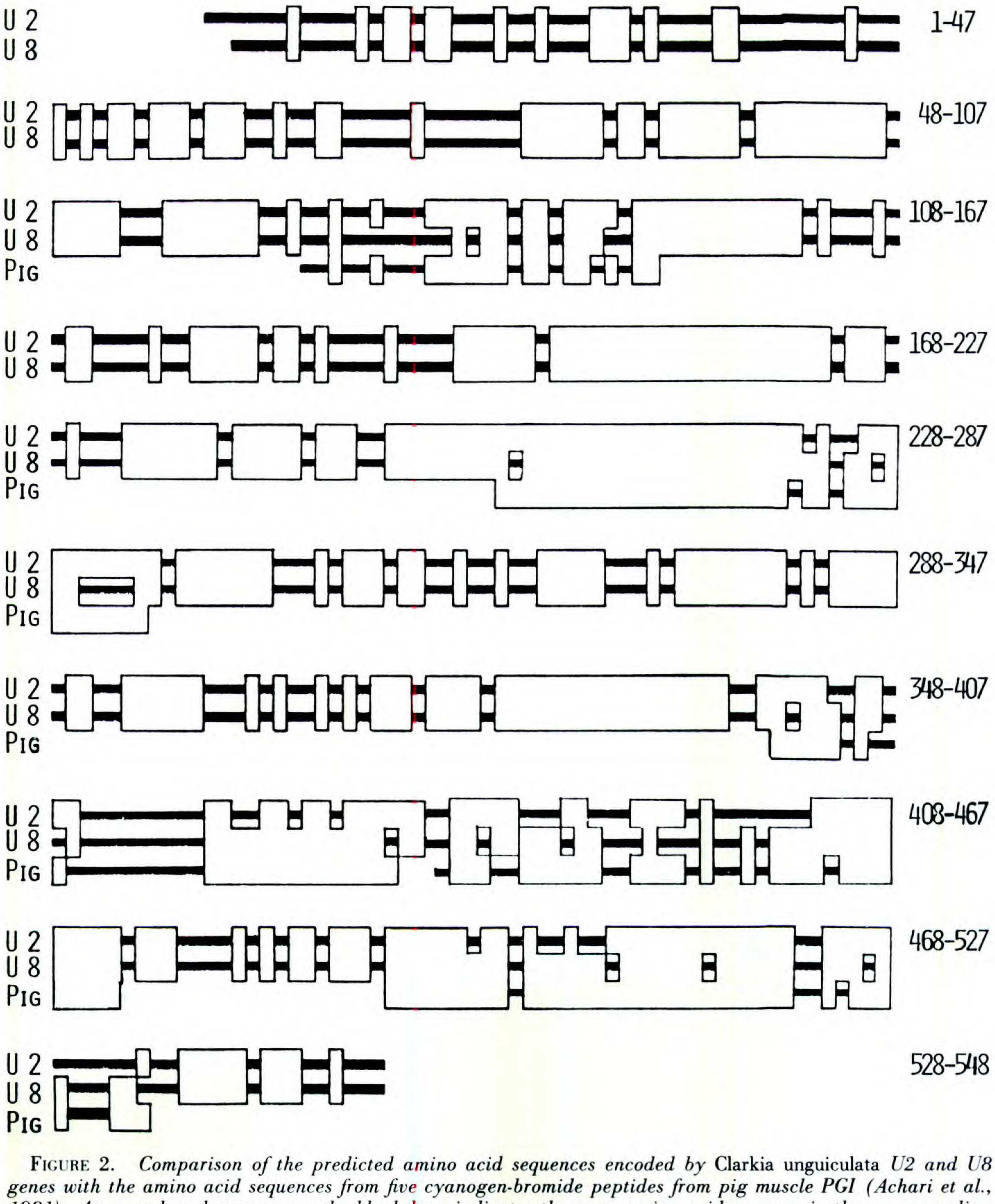

FIGURE 2. *Comparison of the predicted amino acid sequences encoded by* Clarkia unguiculata *U2 and U8 genes with the amino acid sequences from five cyanogen-bromide peptides from pig muscle PGI (Achari et al., 1981). An open box drawn across the black bars indicates the same amino acid appears in the corresponding position on two or three of the sequences. The amino acids are numbered on the right beginning with the first methionine in the U2 sequence. U2 encodes 548 amino acids, U8 encodes 543 amino acids, and the total number of amino acids identified in the pig peptides is 166. The diagram represents the best fit by eye, taking into account several short insertions and deletions in the sequences.*

mology to pig PGI (Table 2), and encodes a protein of 543 amino acids.

The predicted amino acid sequences show that U2 and U8 are 58% homologous over their entire coding regions. Comparing U2 and U8 only in the regions covered by the pig peptides, the two sequences are 54% identical (Table 2). Thus, the two *Clarkia* PGI genes differ more from each other than either does from pig PGI. The homology of the *Clarkia* and pig sequences is diagrammed in Figure 2. The two *Clarkia* proteins exhibit large blocks of very high amino acid identity as well as

many shorter regions of nonidentity. Several lengthy portions of the three sequences show complete identity. Overall, the high homology between the pig PGI amino acid sequences and the predicted *Clarkia* amino acid sequences establishes with certainty that both *Clarkia* genes encode PGI.

On the basis of lack of interruption in their open reading frames and the lengths of their sequences, which encode proteins that have closely similar molecular weights to that previously determined for *Clarkia* PGI, neither *Clarkia* gene appears to include introns (and see below). Otherise, both genes have many features expected of eukaryotic genes, including potential TATA boxes and other upstream regions similar to known regulatory sequences. A complete transcriptional characterization of the genes will be reported separately.

Clarkia unguiculata possesses the PGI duplication, and its genome must include two loci encoding cytosolic PGIs and one locus encoding plastid PGI. Since a heterologous probe was used to obtain the U2 and U8 PGI genes, it was necessary to determine which isozyme is encoded by each gene. A priori, the expectation was that sequences encoding the cytosolic PGIs would be more similar to each other than either would be to the plastid PGI. Genes encoding plastid and cytosolic glycolytic isozymes have been cloned and sequenced in plants only for tobacco glyceraldehyde-3-phosphate dehydrogenase (G3PD) (Shih et al., 1986), and the results of that study are closely relevant to our research with PGI. Comparison of predicted amino acid sequences from cDNAs showed that the tobacco cytosolic G3PD was more similar to other eukaryotic G3PD enzymes, with about 65% homology, than it was to the tobacco plastid isozyme, with 45% homology. The homology of U2, U8, and pig PGI are roughly similar to these values, but we were able to compare only a few sequences.

Our initial attempt to identify the isozymes encoded by the *Clarkia* genes centered on the search for correlation between restriction length fragments and allelic segregation. This could be followed by PGI activity staining on starch gels following electrophoresis of leaf extracts and correlated with the RFLP segregation. To date, we have examined a number of DNAs from single *C. unguiculata* plants by restriction analysis followed by electrophoresis and Southern blotting. The DNAs proved highly polymorphic, but we have been able to match restriction fragments to U2. U8 and several other genes cloned from the *C. unguiculata* library have not yet been similarly matched, but this is not unexpected, since the library was made from DNA isolated from a bulk of hundreds of individual seedlings and this commercial strain is highly polymorphic.

However, a different procedure suggests that U8 encodes the slowly migrating allozyme PGI-3B, a cytosolic isozyme. The U8 sequence was inserted in pUC18, downstream from the beta-galactosidase promoter. When the operon was induced by IPTG, the *E. coli* host synthesized very large quantities of PGI protein. The PGI was catalytically active and had a very slightly faster electrophoretic mobility on starch gels than the slow allozyme PGI-3B of *C. unguiculata*, a difference probably caused by different post-translational protein modification between *Clarkia* and *E. coli*. By the same procedure, a large quantity of protein with the molecular weight of PGI was also synthesized from U2, and its electrophoretic mobility was similar to that of *Clarkia* plastid PGI. The expression of these genomic clones in *E. coli*, apparently by virtue of fortuitous promoters in their 5′ noncoding region, provides convincing evidence that introns are not present in these genes. Whether other PGI genes also lack introns remains to be determined. Their absence is surprising, since other genes encoding glycolytic enzymes in plants such as maize *Tpi* has eight introns (Marchionni & Gilbert, 1986) and maize *Adh* has nine introns (Dennis et al., 1984).

To summarize the molecular studies, we have cloned and sequenced two PGI genes from a genomic library of *C. unguiculata*, a species with the PGI duplication. The genes have a homology of 58%; one of them (U8) appears to encode a cytosolic PGI-3 isozyme; the other is thought to encode a plastid PGI. We have also constructed genomic libraries from *Clarkia* species without the PGI duplication and have obtained clones of a number of sequences homologous to the PGI probes from *C. unguiculata*. The molecular genetics studies of PGI in *Clarkia* constitute one of the first analyses of the evolution of a plant nuclear gene. Many additional molecular studies are called for to understand gene evolution and to improve phylogenetic reconstructions.

Literature Cited

Achari, A., S. E. Marshall, H. Muirhead, R. H. Palmieri & E. A. Noltmann. 1981. Glucose-6-phosphate isomerase. Philos. Trans., Ser. B 293: 145–157.

Burnham, C. H. 1962. Discussions in Cytogenetics. Burgess Publ. Co., Minneapolis, Minnesota.

Constance, L. 1964. Systematic botany: an unending synthesis. Taxon 13: 257–273.

Crawford, D. 1983. Phylogenetic and systematic inferences from electrophoretic studies. Pp. 257–287

in S. O. Tanksley & T. J. Orton (editors), Isozymes in Plant Genetics and Breeding, Part A. Elsevier, Amsterdam.

———. 1985. Electrophoretic data and plant speciation. Syst. Bot. 10: 405–416.

DAVIS, P. H. & V. H. HEYWOOD. 1967. Principles of Angiosperm Taxonomy. Van Nostrand, Princeton, New Jersey.

DAVIS, W. S. 1970. The systematics of *Clarkia bottae*, *C. cylindrica*, and a related new species, *C. rostrata*. Brittonia 22: 270–284.

DENNIS, E. S., W. L. GERLACH, A. J. PRYOR, J. L. BENNETZEN, A. INGLIS, D. LLEWELLYN, M. M. SACHS, R. J. FERL & W. J. PEACOCK. 1984. Molecular analysis of the alcohol dehydrogenase (*Adh1*) gene of maize. Nucl. Acids Res. 12: 3983–4000.

FISCHER, R. L. & R. B. GOLDBERG. 1982. Structure and flanking regions of soybean seed protein genes. Cell 29: 651–660.

GOTTLIEB, L. D. 1973. Enzyme differentiation and phylogeny in *Clarkia franciscana*, *C. rubicunda*, and *C. amoena*. Evolution 27: 205–214.

———. 1974a. Genetic confirmation of the origin of *Clarkia lingulata*. Evolution 28: 244–250.

———. 1974b. Gene duplication and fixed heterozygosity for alcohol dehydrogenase in the diploid plant *Clarkia franciscana*. Proc. Natl. Acad. U.S.A. 71: 1816–1818.

———. 1977. Evidence for duplication and divergence of the structural gene for phosphoglucose isomerase in diploid species of *Clarkia*. Genetics 86: 289–307.

———. 1981. Electrophoretic evidence and plant populations. Progress in Phytochem. 7: 1–46.

———. 1982. Conservation and duplication of isozymes in plants. Science 216: 373–380.

———. 1983. Isozyme number and phylogeny. Pp. 209–221 *in* U. Jensen & D. E. Fairbrothers (editors), Proteins and Nucleic Acids in Plant Systematics. Springer-Verlag, Berlin, Heidelberg.

———. 1986. Genetic differentiation, speciation and phylogeny in *Clarkia* (Onagraceae). Pp. 145–160 *in* K. Iwatsuki, P. H. Raven & W. J. Bock (editors), Modern Aspects of Species. Univ. of Tokyo Press, Tokyo.

——— & R. C. HIGGINS. 1984a. Evidence from subunit molecular weight suggests hybridization was the source of the phosphoglucose isomerase gene duplication in *Clarkia*. Theor. Appl. Genet. 68: 369–373.

——— & ———. 1984b. Phosphoglucose isomerase expression in species of *Clarkia* with and without a duplication of the coding gene. Genetics 107: 131–140.

——— & N. F. WEEDEN. 1979. Gene duplication and phylogeny in *Clarkia*. Evolution 33: 1024–1039.

HENIKOFF, S. 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene 28: 351–359.

HIGGINS, R. C. & L. D. GOTTLIEB. 1984. Subunit hybridization and immunological studies of duplicated phosphoglucose isomerase isozymes. Biochem. Genet. 22: 957–979.

JANSEN, R. K. & J. D. PALMER. 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proc. Natl. Acad. U.S.A. 84: 5818–5822.

JONES, T. W. A., E. PICHERSKY & L. D. GOTTLIEB. 1986. Enzyme activity in EMS-induced null mutations of duplicated genes encoding phosphoglucose isomerases in *Clarkia*. Genetics 113: 101–114.

LEWIS, H. 1953. The mechanism of evolution in the genus *Clarkia*. Evolution 7: 1–20.

———. 1962. Catastrophic selection as a factor in speciation. Evolution 16: 257–271.

———. 1973. The origin of diploid neospecies in *Clarkia*. Amer. Naturalist 107: 161–170.

———. 1980. The mode of evolution in *Clarkia* (Onagraceae). Paper presented at ICSEEB II, Vancouver, British Columbia.

——— & M. E. LEWIS. 1955. The genus *Clarkia*. Univ. Calif. Publ. Bot. 20: 241–392.

——— & P. H. RAVEN. 1958. Rapid evolution in *Clarkia*. Evolution 12: 319–336.

MARCHIONNI, M. & W. GILBERT. 1986. The triosephosphate isomerase gene from maize: introns antedate the plant–animal divergence. Cell 46: 133–141.

MESSING, J. 1983. New M13 vectors for cloning. Meth. Enzymol. 101: 20–78.

ODRZYKOSKI, I. & L. D. GOTTLIEB. 1984. Duplications of genes coding 6-phosphogluconate dehydrogenase in *Clarkia* (Onagraceae) and their phylogenetic implications. Syst. Bot. 9: 479–489.

PICHERSKY, E. & L. D. GOTTLIEB. 1983. Evidence for duplication of the structural genes coding plastid and cytosolic isozymes of triose phosphate isomerase in diploid species of *Clarkia*. Genetics 105: 421–436.

RAVEN, P. H. 1979. A survey of reproductive biology in Onagraceae. New Zealand J. Bot. 17: 575–593.

SCHWARTZ, D. & T. ENDO. 1966. Alcohol dehydrogenase polymorphism in maize: simple and compound loci. Genetics 53: 709–715.

SHIH, M.-C., G. LAZAR & H. M. GOODMAN. 1986. Evidence in favor of the symbiotic origin of chloroplasts: primary structure and evolution of tobacco glyceraldehyde-3-phosphate dehydrogenase. Cell 47: 73–80.

SOLTIS, P. S., D. E. SOLTIS & L. D. GOTTLIEB. 1987. Phosphoglucomutase gene duplications in *Clarkia* (Onagraceae) and their phylogenetic implications. Evolution 41: 667–671.

SYTSMA, K. J. & L. D. GOTTLIEB. 1986a. Chloroplast DNA evolution and phylogenetic relationships in *Clarkia* sect. *Peripetasma* (Onagraceae). Evolution 40: 1248–1261.

——— & ———. 1986b. Chloroplast DNA evidence for the origin of the genus *Heterogaura* from a species of *Clarkia* (Onagraceae). Proc. Natl. Acad. U.S.A. 83: 5554–5557.

TANKSLEY, S. D. 1987. Organization of the nuclear genome in tomato and related diploid species. Amer. Naturalist 130 suppl.: S46–S61.

WEEDEN, N. F. & L. D. GOTTLIEB. 1979. Distinguishing allozymes and isozymes of phosphoglucose isomerases by electrophoretic comparison of pollen and somatic tissues. Biochem. Genet. 17: 287–296.